Elementary Statistical Methods & Tips on Group Project

EC320

September 5, 2014

Roadmap

1. Regression Analysis
   - Understanding the Concept of Regression
   - Interpretation
   - Example
   - Common Problems

2. Tips on Group Project
   - Literature
   - Data
   - Statistical Software
   - Examples

1. Regression Analysis

Regression Analysis

- A statistical tool for the investigation of relationships between two or more variables
- Examples (bivariate):
  - Is poverty headcount ratio for a country (y) negatively related to years of education (x)?
  - What is the relationship between children mortality rate (y) and accessibility to clean water (x)?
  - $y = \alpha + \beta x + e$
  - e is all random factors that affect y other than x
- We will focus on linear regression (additive form)

Interpreting the regression coefficient

- After running a regression, an estimate for β, $\widehat{\beta}$, is obtained.

- $\hat{\beta}$ is the size of influence of x on y
  - The expected change in y given a unit change in x

- Multivariate case
  - $y = \alpha + \beta x + \gamma z + e$
  - $\hat{\beta}$ is the expected change in y given a unit change in x with all values of z constant

Example

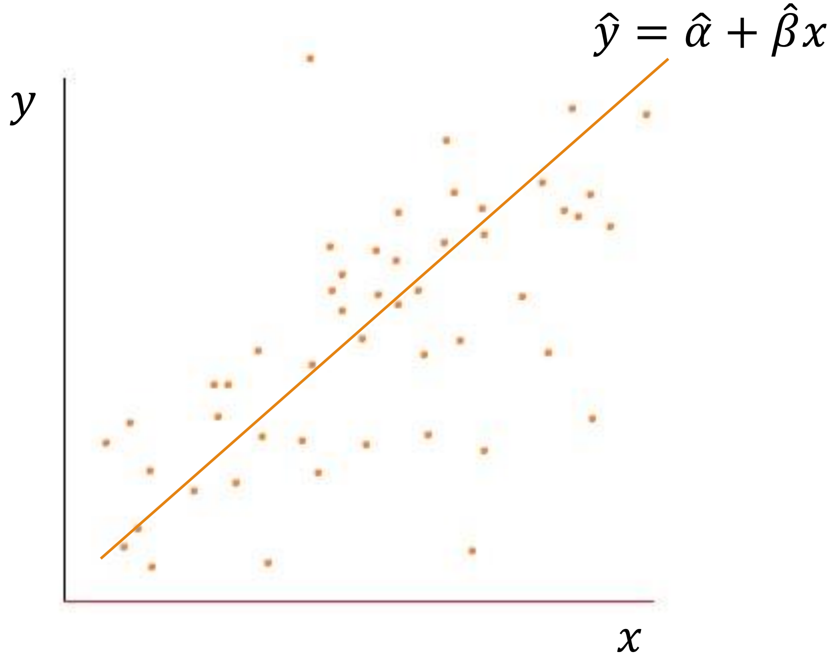
- Consider the factors that contribute to poverty alleviation

- $poverty\ headcount\ ratio = \alpha_1 + \beta_1 * years\ of\ schooling + e$
- $poverty\ headcount\ ratio = \alpha_2 + \beta_2 * years\ of\ schooling + \gamma * per\ capita\ income + u$

- What is the difference between β_1 and β_2?

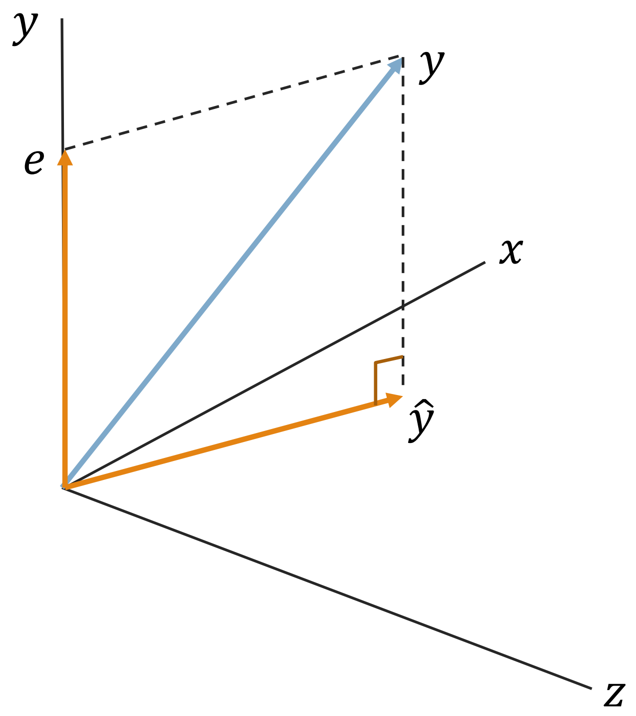
Fitting a Line

Scatter Diagram



- We want to explain y with x
- Set up the model: $y = \alpha + \beta x$
- What is the best line?
- OLS (Ordinary Least Squares)
  - We choose α and β that minimize $\sum_{i=1}^{N}(y - \alpha - \beta x)^2$
  - $y = \hat{y} + e = \hat{\alpha} + \hat{\beta}x + \mathrm{e}$

Fitting a Line



- Adding one explanatory variable

  = Adding one dimension

- What is a reasonable linear combination of x and z that best represents y?

- Set up the model: $y = \beta_1 x + \beta_2 z + e$
- Again, we choose β_1 and β_2 that minimize $\sum_{i=1}^{N}(y_i - \beta_1 x_i - \beta_2 z_i)^2$

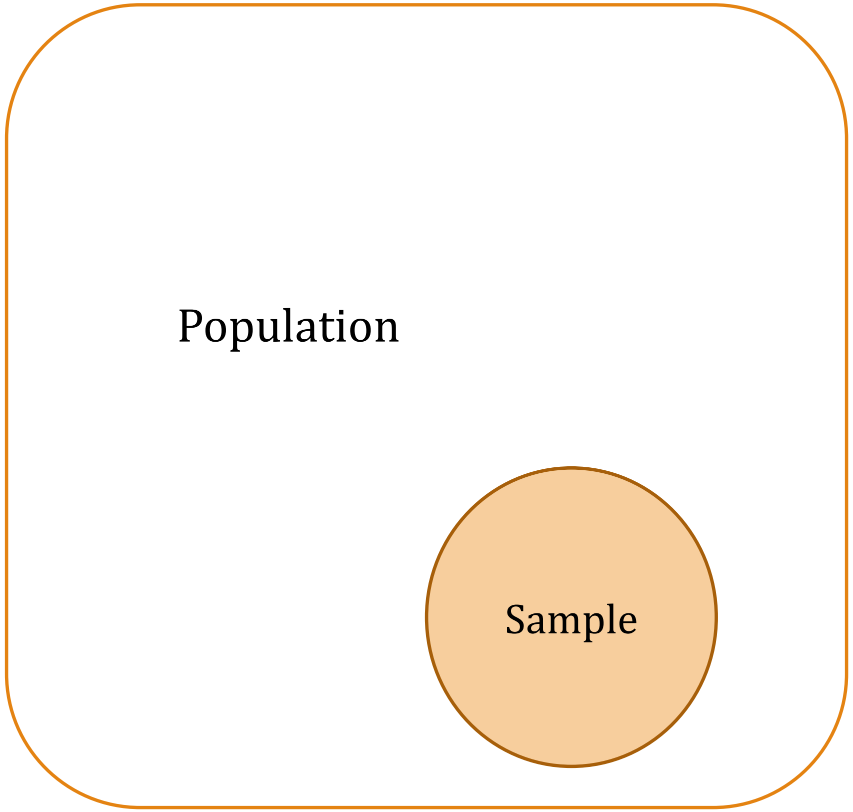
Goodness of fit (R^2)

- $R^2 = \frac{\sum_{i=1}^{N}(\widehat{y_t}-\bar{y})^2}{\sum_{i=1}^{N}(y_t-\bar{y})^2} = \frac{Explained\ Sum\ of\ Squares}{Total\ Sum\ of\ Squares}$

- R^2 is a measure of the explanatory power of the regression
  - Valid only when there is a constant among regressors

- Is a high R^2 always a good sign?

Inference



- Sample vs. population
- The value of $\hat{\beta}$ varies depending on the sample
- We cannot assert that we have found the true value of β by running a regression no matter how large the sample is

$t - ratio$

- $t = \frac{\hat{\beta}}{s.e.}$
- The $t - ratio$ tests whether the estimated coefficient is significantly different from zero
- It follows *t-distribution* with *N-k* degrees of freedom

  (k is the number of independent variables)
- Asymptotically (with a large enough set of observations) it follows *normal distribution*

Significance

- Assuming that the sample is large, we will refer to the normal distribution
- Two-tailed test
- Critical values: 1.645 for 90% significance level
  1.96 for 95% significance level
  2.58 for 99% significance level
- Interpretation

Suppose the t-value is 1.79 and the sample is large enough.

Then the probability that the null ($\beta = 0$) is rejected in favor of the alternative ($\beta \neq 0$) is between 0.9 and 0.95.

$p - value$

- The smallest significance level at which the null (that $\beta = 0$) can be rejected
- If p-value is 0.01, it means that the null can be rejected whenever the size of test is greater than or equal to 0.01

Example

Imai et al. , "Microfinance and Poverty – A Macro Perspective," *World Development*, 2012, pp. 1675–1689

- The model

$$pov_i = \beta_0 + \beta_1 GLP_i + \beta_2 GDPPC_i + \beta_3 DomesticCrd_i + \beta_4 REG_i + u_i$$

  - pov_i: poverty head count ratio
  - GLP_i: gross loan portfolio
  - $GDPPC_i$: gross domestic product per capita at 2000 constant USD prices
  - $DomesticCrd_i$: domestic credit of banks as a proportion of GDP
  - REG_i: regional dummies with Latin America and Caribbean being the reference region

Example

Imai et al. , “Microfinance and Poverty – A Macro Perspective,” *World Development*, 2012, pp. 1675–1689

Explanatory Variables	Pooled OLS	
Log of GLP per capita	-1.31	[-2.53]*
Log of GDP per capita	-9.43	[-5.45]**
Domestic Credit	-0.07	[-2.40]*
2007 Year Dummy	-1.28	[-0.64]
MENA	-10.25	[-3.22]**
EAP	3.74	[0.78]
ECA	-10.96	[-5.32]
SA	14.19	[1.75]
SSA	22.56	[3.73]**
Constant	92.16	[6.67]**
N	99	
Adjusted R square	0.851	
F-statistic	51.81	

- Reported in the parenthesis are *t-values*
- * Significant at 5%
- ** Significant at 1%

Common Problems

- Omitted Variables
  - Omitted variable bias
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$
- Multicollinearity
  - Size of standard error increases
  - If two or more variables are highly correlated, it becomes difficult to separate their explanatory power
- Endogeneity
  - IV (Instrumental Variables)
- Developing countries
  - Missing/unavailable/inaccurate Data

2. Tips on Group Project

How to build a model?

- Thought experiment
- Existing literature

- Plot scatter diagrams
- Make a summary statistics table

Search for Journals and Articles

- BU Library (http://www.bu.edu/library)
  - Access to most published articles with BU ID

- Google Scholar (http://scholar.google.com)
  - Alternative source
  - If an article is not free (even via BU library), try checking 'All versions' of the article

- UNDP and World Bank both publish their own reports.
  - http://www.worldbank.org/en/research
  - http://www.worldbank.org/reference/?lang=en
  - http://www.undp.org/content/undp/en/home/librarypage.html

Search for Data

- World Bank
  - http://data.worldbank.org/
- UN
  - http://data.un.org/
- IMF
  - http://www.imf.org/external/data.htm
- PWT (Penn World Table)
  - https://pwt.sas.upenn.edu/

Types of Data

- Cross-sectional Data
  - Observations at the same point in time but across different units
  - eg., GDP of 60 countries in 2012 (y_i)
- Time-series Data
  - Observations for the same unit but across horizon
  - eg., GDP of the US from 1960 to 2010 (y_t)
- Panel Data
  - Observations across different units and time
  - eg., GDP of G20 from 1960 to 2010 (y_{it})

Types of Variables

- Numerical Variables
  - Continuous variables
    - GDP, stock prices, precipitation, temperature
  - Discrete variables
    - Population, number of hospitals in a town (count variables)

- Categorical Variables
  - Dummy variables
    - Male/female, smoker/non-smoker, employed/unemployed
  - Nominal variables
    - Africa/America/Antarctica/Asia/Australia/Europe

Statistical Software

- Microsoft Excel
  - Cross-sectional or Time-series data
  - Panel data (dummy variables required to capture the fixed effects)
- SPSS
  - Available in the library
- Stata
  - BU students get discount
  - Click-based
- R, Matlab
  - Available in the library
  - Coding needed

Examples

- Cross-sectional
- Time-series
- Two-period panel